Uncertainty Class Activation Map (U-CAM) using Gradient Certainty method

Badri N. Patro[§], Mayank Lunayach, Vinay P. Namboodiri
Indian Institute of Technology, Kanpur

***Abstract*—Understanding and explaining deep learning models is an imperative task. Towards this, we propose a method that obtains gradient-based certainty estimates that also provide visual attention maps. Particularly, we solve for visual question answering task. We incorporate modern probabilistic deep learning methods that we further improve by using the gradients for these estimates. These have two-fold benefits: a) improvement in obtaining the certainty estimates that correlate better with misclassified samples and b) improved attention maps that provide state-of-the-art results in terms of correlation with human attention regions. The improved attention maps result in consistent improvement for various methods for visual question answering. Therefore, the proposed technique can be thought of as a tool for obtaining improved certainty estimates and explanations for deep learning models. We provide detailed empirical analysis for the visual question answering task on all standard benchmarks and comparison with state of the art methods.**

***Index Terms*—Uncertainty, Class Activation Map, Visual Question Answering, Explanation, Attention, CNN, LSTM, Bayesian Model., Aleatoric, Epistemic Uncertainty.**

I. Introduction

To interpret and explain the deep learning models, many approaches have been proposed. One of the approaches uses probabilistic techniques to obtain uncertainty estimates, [1], [2]. Other approaches aim at obtaining visual explanations through methods such as Grad-CAM [3] or by attending to specific regions using hard/soft attention. With the recent probabilistic deep learning techniques by Gal and Ghahramani [1], it became feasible to obtain uncertainty estimates in a computationally efficient manner. This was further extended to data uncertainty and model uncertainty based estimates [4]. Through this work, we focus on using gradients uncertainty losses to improve attention maps while also enhancing the explainability leveraging the Bayesian nature of our approach. The uncertainties that we use are aleatoric and predictive[5]. For the estimated uncertainties, we calculate gradients using the approach similar to gradient-based class activation maps [3]. This provides "certainty maps", which helps in attending to certain regions of the attention maps. Doing this, we report an improvement in attention maps. This is illustrated in the Figure 1.

Our method combines techniques from both the explanation [3] and uncertainty [4] estimation techniques to obtain improved results. We have provided an extensive evaluation. We show that the results obtained for uncertainty estimates show a strong correlation with misclassification, i.e., when the classifier is wrong, the model is usually uncertain. Further, the attention maps provide state of the art correlation with human-annotated attention maps. We also show that on various VQA datasets, our model provides results comparable to the state of the art while significantly improving the performance of baseline methods on which we incorporated our approach. Our method may be seen as a generic way to obtain Bayesian uncertainty estimates, visual explanation, and as a result, improved accuracy for the Visual Question Answering (VQA) task.

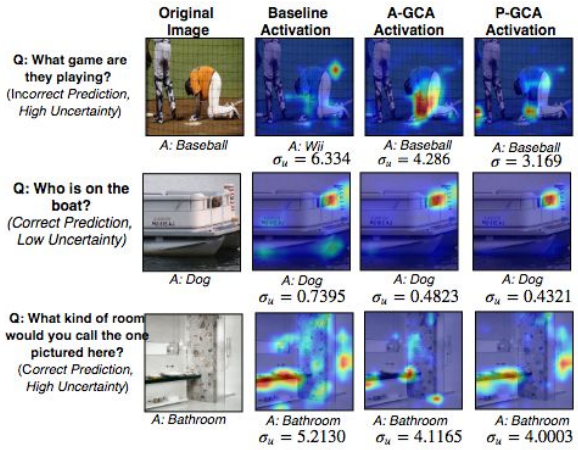

Fig. 1. The figure shows the activation maps for baseline (MCB[6]) and our models (Aleatoric Gradient Certainty for Attention (A-GCA) and Predictive Gradient Certainty for Attention (P-GCA)). In the first example, the baseline model had predicted the wrong answer and had high uncertainty in prediction. (σ_u denotes uncertainty, see Section III). Our model gave a correct answer while also minimizing the uncertainty (thus leading to an improved visual explanation). The source of original images are refereed from MS-COCO [7] dataset and questions are from VQA [8] dataset.

Our contributions, lie in, a) unifying approaches for understanding deep learning methods using uncertainty estimate and explanation b) obtaining visual attention maps that correlate best with human-annotated attention regions and c) showing that the improved attention maps result in consistent improvement in results. This is particularly suited for vision and language-based tasks where we are interested in understanding visual grounding, i.e., for instance, if the answer for a question is 'dog' (Corresponding question: 'Who is on the boat?'), it is important to understand whether the model is certain and whether it is focusing on the correct regions containing a dog. This important requirement is met by the proposed approach.

In our previous work [27], we proposed a gradient-based uncertainty explanation method to provide visual explanation

[§]Currently working at Google

TABLE I
OVERVIEW OF ATTENTION AND EXPLANATION METHOD FOR DEEP NETWORKS. (IG: INTEGRATED GRADIENT, GBP: GUIDED BACK-PROPAGATION, WTQ: WIKITABLE QUESTION, IN: IMAGENET, CO: MS-COCO, TAL: TAL-GATA)

Methods	**Base Model**	**Improving Attention**	**Improving Explanation**	**Visualisation**	**Dataset**
Deconv[9]	discriminative	✗	✓	imshow	MNIST[10]
GBP [11]	discriminative	✗	✓	imshow	CIFAR[12], IN[13]
DeepLift [14]	discriminative	✗	✓	imshow	MNIST[10], TAL[15]
IG [16]	discriminative,generative	✗	✓	imshow	IN[13], DR[17], WTQ[18]
Grad-CAM [3]	discriminative, generative	✗	✓	camshow	CO[7], VQA[8], VOC[19]
RISE [20]	discriminative, generative	✗	✓	camshow	CO[7], VOC[19]
VQA [8]	discriminative	✗	✗	attention map	VQAv1 [8]
SAN [21]	discriminative	✓	✗	attention map	VQAv1 [8]
MCB-att[6]	discriminative	✓	✗	attention map	VQAv1 [8]
DAN [22]	discriminative	✓	✗	attention map	VQAv1 [8], VQAv2 [23]
BAN [24]	discriminative	✓	✗	attention map	VQAv1 [8]
PUDN [25]	generative	✓	✗	attention map	Visual Dialog[26]
U-CAM (Ours)	discriminative	✓	✓	attention map, camshow	VQAv1 [8], VQAv2 [23]

for the task of Visual Question Answering (VQA). We obtain gradient certainty map for attention in the VQA task. In this work, we provide further analysis of the experimental results, detailed analysis of the role played by uncertainty without attention and uncertainty with attention. We also justify the need for multi-modal embedding for obtaining the uncertainty based attention. We further provide additional analysis for the estimated uncertainty and visualise the uncertainty present in the model. This is considered in detail in the main manuscript.

For contextualizing our work, we provide a comparison of the various models proposed in the literature in the Table I. We report various visualizations that the previous works use. "Imshow" visualizes explanation maps in the image space, while "camshow" visualizes explanation maps in the feature space using the class activation map technique. This table also provides information about the model types (Generative and Discriminative), dataset/tasks(Visual Question Answering (VQA), Visual Dialog, Image captioning, and classification tasks) for the corresponding methods and whether these methods improve attention or explanation or both.

II. RELATED WORK

Visual Question Answering. The task of Visual Question Answering [28], [8], [29], [23], [30] is well studied in the vision and language community. [31] and [32] have proposed interesting methods for improving attention in the question. Work that explores image and question jointly and is based on hierarchical co-attention is [33]. An interesting work obtains a varied set of modules for answering questions of different types [34]. [6], [35], [24] advocate multimodal pooling and obtain close to state of the art results in VQA. [22] has proposed an exemplar-based explanation method to improve attention in VQA. We can do systematic comparison of image-based attention while correlating with human annotated attention maps, as shown by [36]. [37], [21], [38], [22] have proposed attention-based methods to highlight important region in an image based the query question for VQA. These methods use questions to attend over specific regions in an image.

Explainability. Early textual explanation models spanned a variety of applications such as medicine [39], feedback for teaching [40] and were generally template-based. We start with a basic explanation method to find discriminative visual patches [41], [42], whereas others aim to understand intermediate features which are important for the end decisions. Zeiler *et al.* [9] proposed a Deconvolutional Network architecture to visualize convolutional layers. CAM [43] provides discriminative image regions used by the model to predict the particular class. Salience map[44], Guided Backpropagation [45], Integrated gradients [16] provides gradient-based method to provide regions that are highly activated by the networks where the strongest neurons are fired. Slight changes in these high activation part affect the output most. Selvaraju *et al.* [3] aims to provide visual explanation for answer prediction. The authors use a gradient based method to obtain the class activation map.

Uncertainty estimation. Blundell *et al.* [46] has proposed a method to learn uncertainty in the weights of the neural network. Kendall *et.al.* [47] has proposed method to measure model uncertainty for image segmentation task. They observed that the softmax probability function approximates relative probability between the class labels, but does not provide information about the model's uncertainty. The work by [1], [48] estimates model uncertainty of the deep neural networks (CNNs, RNNs) with the help of dropout [49]. Teye *et al.* [50] has estimated uncertainty for batch normalized deep networks. The authors in the relevant works [4], [5], [51] have mainly decomposed predictive uncertainty into two major types, namely aleatoric and epistemic uncertainty, which capture uncertainty present in the data and uncertainty about in model respectively. [52] suggested a method to measure predictive uncertainty with the help of model and data uncertainty. Recently, [53] proposed a method to bring two data distributions close for the domain adaption task.

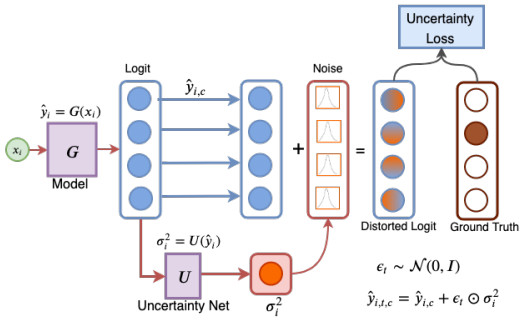

Fig. 2. Illustration of Uncertainty Loss

Alshawi *et al.* [54] have proposed a method to obtain true uncertainty map and estimated uncertainty map from salience map. Salience based method are helping to obtain important regions in case of image and video based system. However, it fails to provide good attention map for multi-model conversation systems like Visual Question Answering and Visual Dialog, as it focuses on more prominent part of the image or video. It does not take account for the conditional query part of the other modality (Text). In VQA, salience based methods do not focus on correct regions [36], while answering the query question based on the image. Attention network [33], [21], [6] helps to obtain important regions in an image with respect to question. We have adopted a concept mentioned in Gal *et al.* [55] to obtain aleatoric uncertainty. While obtaining aleatoric uncertainty we do not have access to the true signal corresponding to the output prediction (the true 'data estimate' corresponding to the output prediction). In order to estimate it, we therefore need to use a surrogate method as suggested in [55] to obtain the corresponding estimate as described in the paper.

In this work, the objective is to analyze and minimize the uncertainty in attention masks for predicting answers in VQA. To accomplish this, we are proposing gradient-based certainty explanation masks, which minimize the uncertainty in attention regions for improving the correct answer's predicted probability in VQA. Our method also provides a visual explanation based on uncertainty class activation maps, capturing and visualizing the uncertainties present in the attention maps in VQA.

III. Modeling Uncertainty

We consider two type of uncertainties to model the uncertainty present in the network, one due to uncertainty present in the data (Aleatoric), and the other due to the model (Epistemic uncertainty).

A. Modeling Aleatoric Uncertainty

Given an input x_i the model (G) predicts the logit output $\hat{y}_i$ which is then an input to uncertainty network (U) for obtaining the variance σ_i^2 as shown in Figure 2. To capture Aleatoric uncertainty [4], we learn the observational noise parameter σ_i for each input point x_i. Then, Aleatoric uncertainty, $(\sigma_a^2)_i$ is estimated by applying softplus function on the output logit variance. This is given by,

$$(\sigma_a^2)_i = Softplus(\sigma_i^2) = log(1 + exp(\sigma_i^2)) \tag{1}$$

For calculating the aleotoric uncertainty loss, we perturb the logit value (y_i) with Gaussian noise of variance $(\sigma_a^2)_i$ (diagonal matrix with one element corrosponding to each logits value) before the softmax layer. The logits reparameterization trick [5] and [55] combines $\hat{y}_{i,c}$ and σ_i to give $\mathcal{N}(\hat{y}_{i,c}, \sigma_i^2)$. We then obtain a loss with respect to ground truth. It is expressed as:

$$\hat{y}_{i,c,t} = y_{i,c} + \epsilon_t * \sigma_i^2, where \quad \epsilon_t \sim \mathcal{N}(0, I) \tag{2}$$

$$\mathcal{L}_a = \sum_i \log \frac{1}{T} \sum_t \exp\left(\hat{y}_{i,c,t} - \log \sum_{c^{'}} \exp \hat{y}_{i,c^{'},t}\right) \tag{3}$$

where $\mathcal{L}_a$ is the aleatoric uncertainty loss (AUL), T is the number of Monte Carlo simulations. $c^{'}$ is a the class index of the logit vector $y_{i,t}$ which is defined for all the classes.

B. Modeling Predictive Uncertainty

To obtain the model uncertainty, we measure epistemic uncertainty. However, estimating epistemic uncertainty[52] is computationally expensive, and thus we measure the predictive uncertainty, having both aleatoric and epistemic uncertainties present in it. To estimate it, we sample weights in the Bayesian networks G and then perform Monte Carlo simulations over the model to obtain the predicted class probabilities $p(y_{i,t})$. That is,

$$O(\hat{y}_{i,t}) = G^t(x_i) \quad v_{i,t}^a = Softplus(U^t(\hat{y}_{i,t}))$$
$$p(\hat{y}_{i,c}|x_i, X_I) = (\frac{1}{T}\sum_{t=1}^{T} SoftmaxO(\hat{y}_{i,t}))_c$$

where c is the answer class, $G^t \sim G, U^t \sim U$ and $v_{i,t}^a$ is the aleatoric variance of each logit in the tth MC Simulation. The entropy of the sampled logit's probabilities can be calculated as:

$$H(\hat{y}_i) = -\sum_{c=1}^{C} p(\hat{y}_{i,c}) * \log p(\hat{y}_{i,c}) \tag{4}$$

The predictive uncertainty contains entropy and aleatoric variance when it's expectation is taken across T number of Monte Carlo simulations:

$$\sigma_p^2 = H(\hat{y}_i) + \frac{1}{T}\sum_{t=1}^{T} v_{i,t}^a \tag{5}$$

where $H(\hat{y}_i)$ is the entropy of the probability $p(\hat{y}_i)$, which depends on the spread of class probabilities while the variance (second term in the above equation) captures both the spread and the magnitude of logit outputs, $\hat{y}_{i,t}$. In Equation 2, we can replace σ_a^2 with predictive uncertainty σ_p^2 (mentioned above in Equation 5) to get the predictive uncertainty loss (PUL).

C. Attention, Explanation and Uncertainty

We provide the connection between uncertainty, explanation, and attention to explain these better.

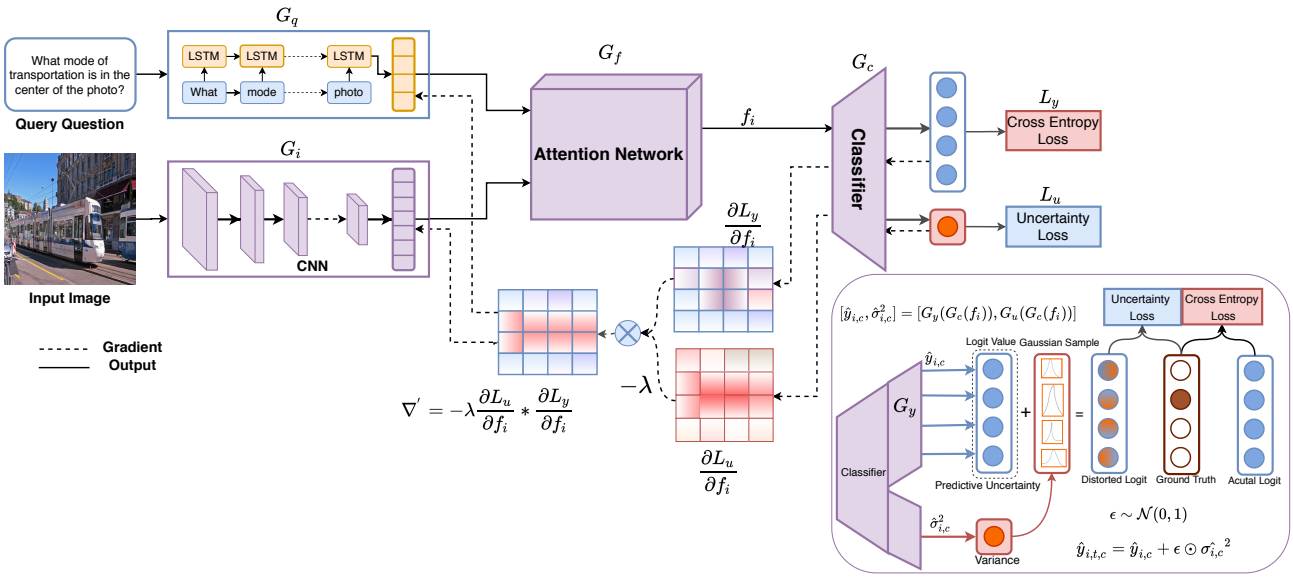

Fig. 3. Illustration of our proposed method Gradient Certainty for Attention Mask (GCA) and its certainty mask. We obtain image feature and question feature using CNN and LSTM, respectively. We then obtain attention mask using these features, and classification of the answer is done based on the attended feature.

Attention: Attention is the weighted sum of the spatial regions in the input. We need to learn those weights of the attention network with the help of back propagation. Model learns to assign a probability score on the various regions of an image. In VQA, an attention network learns to focus on important regions in an image with respect to the given query Question which helps in reaching to a correct answer.

Explanation: The main distinction between attention and explanation is that the explanation is conditioned on the answer. That is, we know that given a particular label, what evidence do we have with respect to the activation map that helps explain why this conclusion was reached. This is obtained by back-propagating with respect to the true label and it provides us the important regions in the image because of which the decision was made by the model. Both attention and explanation are not same. Attention learns the weights to provide correct responses, while explanation uses the learned weights to provide reasons, i.e. why the decision was made.

Uncertainty: It provides, how much the model is confused while predicting the responses or how confused the input is. Using these scores, we obtain a mask for our attention network. This mask informs which portion in the image are confused and which portions are not. We use the weight combination of this uncertainty to train our model for improving confidence in predictions. It is the confusion present in the input data or in the model while predicting the responses.

IV. Method

Task: We solve for VQA task. The key difference in our architecture as compared to the existing VQA models is the introduction of gradient-based certainty maps. A detailed figure of the model is given in the Figure 3. We keep other aspects of the VQA model unchanged. In a typical open-ended VQA task, we have a multi-class classification task. A combined (image and question) input embedding is fed to the model. Then, the output logits are fed to a softmax function, giving probabilities of the predictions in the multiple-choice answer space. That is, $\hat{A} = \underset{A \in \Omega}{argmax} P(A|I, Q, \theta)$, where Ω is a set of all possible answers, I is the image, Q is the corresponding question, and θ is representing the parameters of the network.

A. Brief about the proposed method

Here we describe more precisely, the role of attention and uncertainty for the VQA task.

Our model takes an image and a question as the input and obtains a logit score from the classifier. The uncertainty is obtained by perturbing the output logits. Let us compare two output predictions, once when the output vector is [1, 0 , 0] and the other case when it's [0.35, 0.32, 0.33]. This is, for a 3 class classifier. In the second case, the output will result in a substantial gradient when perturbed with some noise. This is because the two classes are close to each other. The gradient in the second case, with respect to attended regions will enable us to obtain a gradient update that will modify the attention region to make the output more confident and certain. This is how the uncertainty plays a major role while predicting answer for VQA task.

We perturb the logits and obtain an uncertainty gradient of distorted logits as shown in the Figure 3. We obtain its gradient with respect to the feature f_i (which is an attended feature) and we also obtain a gradient of the actual predicted output with respect to f_i.

Thus, there are two gradients, one with respect to the estimated uncertainty and the other with respect to the output. As these are with respect to f_i (where f_i is the attended feature), we can get a gradient with respect to the attention

mask for the output prediction as well as the uncertainty estimate.

B. U-CAM Approach

The three main parts of our method are Attention Representation, Uncertainties Estimation, and computing gradients of uncertainty losses. In the following sections, we explain them in detail.

1) Attention Representation: We obtain an embedding, $g_i \in \mathcal{R}^{u \times v \times C}$ where u is width, v is height of the image and C represents the number of applied filters on the image X_i in the convolution neural network (CNN). The CNN is parameterized by a function $G_i(X_i, \theta_i)$, where θ_i represents the weights. Similarly, for the query question X_Q, we obtain a question feature embedding g_q using a LSTM network. This network is parameterized by a function $G_q(X_q, \theta_q)$, where θ_q represents the weights. Both g_i and g_q are fed to an attention network that combines the image and question embeddings using a weighted softmax function and produces a weighted output attention vector, g_f as illustrated in Figure 3. People have tried with the various kinds of attention networks. In this paper, we tried with SAN [21] and MCB [6]. Finally, we obtain attention feature f_i using attention extractor network G_f : $f_i = G_f(g_i, g_q)$. The attended feature f_i is passed through a classifier and the model is trained using the cross-entropy loss. Many a times, model is not certain about the answer class to which the input belongs, which sometimes leads to decrease in accuracy. To tackle this, we have proposed a technique to reduce the class uncertainty by increasing the certainty of the attention mask. Additionally, we also incorporate a loss based on the uncertainty which is described next.

2) Estimating Uncertainties: Aleatoric & Predictive: The attention feature, f_i obtained from the previous step is fed to the classifier network(G_c). The output of the classifier is fed to logit network(G_y), which produces class probabilities, y_i. G_c's output is also fed to a variance predictor network, G_u, which outputs the logits' variance, σ_i as mentioned in the Equation 1. For calculating the aleatoric uncertainty loss, we perturb the logit value (y_i) with Gaussian noise of variance $(\sigma_a^2)_i$ before the softmax layer. The Gaussian likelihood for classification is given by $p(y_i|f_i, w) = \mathcal{N}(y_i; G_y(G_c(f_i; w)), \tau^{-1}(f_i; w))$, where w represents model's parameters, τ is the precision, f_i is the attended fused input, and $G_y(G_c(.))$ is the output logit producing network as shown in the Figure 3 . The above setting represents the perturbation of model output with the variance of the observed noise, τ^{-1}. We make sure that $\tau(.)$ is a positive or positive definite matrix (in case of Multivariate) by using the logit reparameterization trick [5] and [55]. Finally, we then obtain an aleatoric loss, $\mathcal{L}_a$ with respect to ground truth as mentioned in the Equation 3. Our proposed model, which uses this loss as one of the components of its uncertainty loss, is called Aleatoric-GCA (A-GCA). Along with aleatoric loss $\mathcal{L}_a$, we combine $\mathcal{L}_{VE}$ and $\mathcal{L}_{UDL}$ as mentioned in the Equation 10 and 11 respectively to get total uncertainty loss $\mathcal{L}_u$. The classifier is trained by jointly minimizing both the classification loss, $\mathcal{L}_y$ and the uncertainty loss, $\mathcal{L}_u$. In Equation 2, we can replace σ_a^2 with predictive uncertainty σ_p^2 (mentioned above in Equation 5) to get the predictive uncertainty loss(PUL). Accordingly, the model which uses this loss as one of the constituents of its uncertainty loss is called Predictive-GCA (P-GCA). Next, we compute the gradients of standard classification loss and uncertainty loss with respect to attended image feature, f_i. Besides training, we also use these gradients to obtain visualizations describing important regions responsible for answer prediction, as mentioned in the qualitative analysis section. (Section VII-C)

Algorithm 1 Gradient Certainty for Attention (GCA)

procedure GCA(I, Q)
  Input: Image X_I, Question X_Q
  Output: Answer y_c
  while loop **do**
    Attention features $G_f(G_i(X_I), G_q(X_Q)) \leftarrow f_i$
    Answer Logit $G_y(G_c(f_i)) \leftarrow \hat{y}$
    Data Uncertainty $G_v(G_c(f_i)) \leftarrow \sigma_A^2$
    if A-GCA **then**:
      $\sigma_W^2 = \sigma_A^2$
    else if P-GCA **then**:
      $\sigma_W^2 = \sigma_A^2 + H(\hat{y}_{i,t})$, (Ref: eq- 5)
    end if
    Ans cross entropy $\mathcal{L}_y \leftarrow loss(\hat{y}, y)$
    Variance Equalizer $\mathcal{L}_{VE} := \sum ReLU(\exp^{\sigma_w^2} - \exp^I)$,
    while $t = 1 : \#MC - Samples$ **do**
      Sample $\epsilon_t^w \sim \mathcal{N}(0, \sigma_W^2)$
      Distorted Logits: $\hat{y}_{i,t} = \epsilon_t^w + \hat{y}_i$
      Gaussian Cross Entropy $L_p = -\sum y \log p(\hat{y_d}|F(.))$
      Distorted Loss : $\mathcal{L}_{UDL} = \exp(\mathcal{L}_y - \mathcal{L}_p)^2$
      Aleatoric uncertainty loss $\mathcal{L}_u = \mathcal{L}_p + \mathcal{L}_{VE} + \mathcal{L}_{UDL}$
    end while
    Compute Gradients w.r.t f_i, $\nabla_y = \frac{\partial \mathcal{L}_y}{\partial f_i}$, $\nabla_u = \frac{\partial \mathcal{L}_u}{\partial f_i}$
    Certainty Gradients $\nabla_u^{'} = -\lambda \nabla_u * \nabla_y$
    Certainty Activation $\nabla_u^{''} = ReLU(\nabla_u^{'}) + \gamma ReLU(-\nabla_u^{'})$
    Final Certainty Gradients $\nabla_u^{'''} = softmax(\nabla_u^{''})$
    Final Attention Gradient $\nabla_y = \nabla_y + \nabla_u^{'''}$
    update $\theta_f \leftarrow \theta_f - \eta \nabla_y$
  end while
end procedure

3) Gradient Certainty Explanation for Attention: Uncertainty present in the attention maps often leads to uncertainty in the predictions and can be attributed to the noise in data and the uncertainty present in the model itself. We improve the certainty in these cases by adding the certainty gradients to the existing Standard Cross-Entropy (SCE) loss gradients for training the model during backpropagation.

Our objective is to improve the model's attention in the regions where the classifier is more certain. The classifier will perform better by focusing more on certain attention regions, as those regions are more suited for the classification task. We can get an explanation for the classifier output as done in the existing Grad-CAM approach ($\frac{\partial \mathcal{L}_y}{\partial f_i}$). But that explanation does

not take the model and data uncertainties into the account. We improve this explanation using the certainty gradients ($-\frac{\partial \mathcal{L}_u}{\partial f_i}$). If we can minimize uncertainty in the VQA explanation, then uncertainties in the image and question features, and thus uncertainties in the attention regions would be subsequently reduced. It is the uncertain regions which are a primary source for errors in the prediction, as shown in Figure 1.

In our proposed method, we compute the gradient of the Standard Classification (cross entropy) loss L_y with respect to attention feature i.e. $\frac{\partial L_y}{\partial f_i}$ and and also the gradient of the uncertainty loss $\mathcal{L}_u$ i.e. $\frac{\partial \mathcal{L}_u}{\partial f_i}$. The obtained uncertainty gradients are passed through a gradient reversal layer, giving us the certainty gradients, i.e., $-\frac{\partial \mathcal{L}_u}{\partial f_i}$.

$$\nabla_y^{'} = -\lambda \frac{\partial \mathcal{L}_u}{\partial f_i} * \frac{\partial \mathcal{L}_y}{\partial f_i} \tag{6}$$

The positive sign of gradient $\nabla_y^{'}$ indicates that the attention certainty is activated on these regions and vice-versa.

$$\nabla_y^{''} = ReLU(\nabla_y^{'}) + \gamma ReLU(-\nabla_y^{'}) \tag{7}$$

We apply a ReLU activation function on the product of gradients of the attention map and the gradients of certainty as we are only interested in attention regions that have a positive influence on interested answer class, i.e. attention regions whose intensity should be increased in order to increase answer class probability y_c, whereas negative values are multiplied by γ (large negative number) as the negative attention regions are likely to belong to other categories in image. As expected, without this ReLU, localization maps sometimes highlights more than just the desired class and achieves lower localization performance. Then we normalize $\nabla_y^{''}$ to get attention regions which are highly activated and giving more weight to certain regions.

$$\nabla_y^{'''} = \frac{(\nabla_y^{''})_{u,v}}{\sum_u \sum_v (\nabla_y^{''})_{uv}} \tag{8}$$

Images with higher uncertainty are equivalent to having lower certainty, so the certain regions of these images should have lower attention values. We use residual gradient connection to obtain the final gradient, which is the sum of gradient mask of $\mathcal{L}_y$ (with respect to attention feature) and the gradient certainty mask $\nabla_y^{'''}$ and is given by:

$$\frac{\partial L_y}{\partial f_i} = \frac{\partial L_y}{\partial f_i} + \nabla_y^{'''} \tag{9}$$

Where $\frac{\partial L_y}{\partial f_i}$ is the gradient mask of $\mathcal{L}_y$ when gradients are taken with respect to attention feature. More details are given in the Algorithm 1.

C. Cost Function

We estimate aleatoric uncertainty in logits space by perturbing each logit using the variance obtained from data. The uncertainty present in the logits value can be minimized using cross-entropy loss on Gaussian distorted logits, as shown in the Equation 3. The distorted logit is obtained using a Gaussian multivariate function, having positive diagonal variance. To stabilize the training process [55], we add an additional term to the uncertainty loss, calling it Variance Equalizer(VE) loss, $\mathcal{L}_{VE}$.

$$\mathcal{L}_{VE} = \exp(\sigma_i^2) - \exp({\sigma_0}^2) \tag{10}$$

where σ_0 is a constant. The uncertainty distorted loss (UDL) is the difference between the typical cross-entropy loss and the aleatoric/predictive loss estimated in the Equation 3. The scalar difference is passed to an activation function to enhance the difference in either direction and is given by :

$$\mathcal{L}_{UDL} = \begin{cases} \alpha(\exp^{[\mathcal{L}_p - \mathcal{L}_y]} - 1), & \text{if } [\mathcal{L}_p - \mathcal{L}_y] < 0. \\ [\mathcal{L}_p - \mathcal{L}_y], & \text{otherwise.} \end{cases} \tag{11}$$

By putting this constraint, we ensure that the predictive uncertainty loss does not deviate much from the actual cross-entropy loss. The total uncertainty loss is the combination of Aleatoric (or prediction uncertainty loss), Uncertainty Distorted Loss, and Variance equalizer loss.

$$\mathcal{L}_u = \mathcal{L}_p + \mathcal{L}_{VE} + \mathcal{L}_{UDL} \tag{12}$$

The final cost function for the network combines the loss obtained through uncertainty (aleatoric or predictive) loss $\mathcal{L}_u$ for the attention network with the cross-entropy.

The cost function used for obtaining the parameters θ_f of the attention network, θ_c of the classification network, θ_y of the prediction network and θ_u for uncertainty network is as follows:

$$C(\theta_f, \theta_c, \theta_y, \theta_u) = \frac{1}{n} \sum_{j=1}^{n} L_y^j(\theta_f, \theta_c, \theta_y) + \eta L_u^j(\theta_f, \theta_c, \theta_u)$$

where n is the number of examples, and η is the hyper-parameter which is fine-tuned using validation set, L_y is standard cross-entropy loss and L_u is the uncertainty loss. We train the model with this cost function until it converges so that the parameters. $(\hat{\theta}_f, \hat{\theta}_c, \hat{\theta}_y, \hat{\theta}_u)$ deliver a saddle point function

$$(\hat{\theta}_f, \hat{\theta}_c, \hat{\theta}_y, \hat{\theta}_u) = \arg \max_{\theta_f, \theta_c, \theta_y, \theta_u} (C(\theta_f, \theta_c, \theta_y, \theta_u)) \tag{13}$$

V. Generalisation of Uncertainty-CAM for VQA system

We observe that uncertainty in the answer prediction is too high, when we have noise in the input data or in the model. For example, in the base line model if you add slight noise in either the input image or question, the baseline VQA model predicts wrong output, that is the answer may change from "Napkin" to "Paper" or "Wii" to "Tennis" as shown in the Figure 4. We minimise uncertainty in the answer prediction using Gradient certainty method as mentioned in the section-IV-B and obtain result as shown in right side of the Figure 4. As can be observed, the predictions for our model are not only correct, they are also more robust with the second (wrong) answer having far lesser scores. In Table-II, we start with a simple VQA model (LSTM Q+I)[8], in which we obtain image feature using a pretrained CNN model (VGG-16) and a question feature using standard LSTM network. We then use a fused network to bring these two embedding

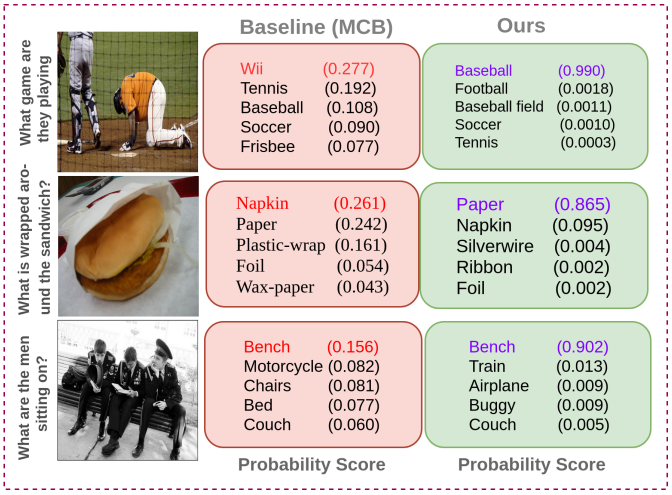

Fig. 4. The figure shows Top-5 answer probability scores for three cases. For the first two rows, baseline predicts wrong answers, and it is confusing (the lesser difference between probabilities of Top 2 predicted classes). In the third example, it (our model) correctly predicts the answer, but its output difference of probability between the top two classes is also very high when compared to the baseline model. The source of original images are refereed from MS-COCO [7] dataset and questions are from VQA [8] dataset.

TABLE II
CONTRIBUTION OF UNCERTAINTY IN VQA

Models	All	Y/N	Num	Oth
LSTM Q+I[8]	53.7	78.9	35.2	36.4
SAN [21]	58.7	79.3	36.6	46.1
LSTM Q+I+Uncertainty	54.1	79.5	35.3	37.0
SAN+Uncertainty	59.5	80.1	36.4	46.7
SAN + P-GCA (ours)	60.4	80.7	36.8	47.9

features close to each other and predict the answer. The second method we tried was with stacked attention network (SAN) [21]. We train our VQA model using uncertainty loss and we observe that there is an increase in accuracy in VQA. We further improve our model accuracy and attention map using Gradient certainty based attention map mechanism. The results are shown in the last row of the Table-II. So in this work, we thoroughly analyse uncertainty-CAM with attention and without attention for VQA models. The observations that we can draw from our analysis are as follows: a) A baseline VQA model (without attention) improves marginally (0.4%) by incorporating uncertainty minimization loss. b) The improvement on a VQA model with attention (SAN) by using uncertainty is more significant (0.8%). c) Jointly considering attention and uncertainty through our proposed model on a stacked attention network (SAN) model results in further improvement. The resulting improvement is 1.7%.

This analysis suggests that using uncertainty results in a consistent marginal improvement. This is more once we have an attention module. However, using joint attention and uncertainty results in the most improvement in the results. Moreover, through this approach our interpretability of the model also improves further as the resultant attention map is more appropriate.

VI. DATA UNCERTAINTY IN A MULTI-MODAL SETTING

Uncertainty in VQA task is two-fold. Consider the example shown in the Figure 5. In the example, the question is, "*Which kind of animal is it?*". When this question is asked (irrespective of image), it may not be concretely answered. Also, seeing the image alone, in the given setting, the animal (especially the one behind) could easily be mis-classified as a dog or some other animal. Specifically, when we attend on image alone, we observe that the attention is only on the animal behind that could be confused with a dog, unlike the animal in the front. These kinds of data uncertainties are tapped & hence minimized best when we consider uncertainties of the fused input (image+question). We tried with minimizing uncertainties of image and question inputs alone. Minimizing the uncertainties in the combined image-question embedding helps in ensuring that more certain regions (and hence better attention maps) are attended to. In Figure 5, we show the resultant attention maps of baseline (when not minimizing uncertainty) & when we tried to minimize only-image, only-question & the combined uncertainty respectively. The architectures used for obtaining these inputs are provided in the Figure 6. The architecture that considers the fused image and question input utilizes all the information and is therefore able to obtain the attention map as shown in the example in Figure 5. We can conclude that by considering both image and question as a input to VQA model, we get better in certainty map and corresponding attention map as compared to only image or or only question as an input.

VII. EXPERIMENTS

We evaluate the proposed GCA methods and have provided both quantitative analysis and qualitative analysis. The former includes: i) Ablation analysis of proposed models (Section-VII-B1), ii) Analysis of uncertainty effect on answer predictions (Figure 7 (a,b)), iii) Differences of Top-2 softmax scores for answers for some representative questions (Figure 7 (c,d)) and iv) Comparison of attention map of our proposed uncertainty model against other variants using Rank correlation (RC) and Earth Mover Distance (EMD) [56] as shown in Table-V for VQA-HAT [36] and in Table- IV for VQA-X [57] . Finally, we compare PGCA with state of the art methods as mentioned in Section-VII-D . Qualitative analysis includes visualization of certainty activation maps for some representative images as we move from our basic model to the P-GCA model. (Section VII-C)

A. Datasets

VQA-v1 [8]: We conduct our experiments on VQA benchmark VQA-v1 [8] dataset, which contains human-annotated questions and answers based on images on MS-COCO dataset. This dataset includes 2,04,721 images in total, out of which 82,783 images are for training, 40,504 images for validation, and 81,434 images for testing. Each image is associated with three questions, and each question has ten possible answers. There are 248349 Question-Answer pairs for training, 121512 pairs for validation, and 244302 pairs for testing.

VQA-v2[23]: We provide benchmark result on VQA-v2[23] dataset. This dataset removes bias present in VQA-v1 by

Fig. 5. Column (a) shows the original image. Columns (b), (c), (d), and (e) represent the baseline attention, attention when only image uncertainty was minimized, attention when only question uncertainty was minimized, attention when both image and question uncertainties were minimized (proposed model), respectively. The source of original images are refereed from MS-COCO [7] dataset and questions are from VQA [8] dataset.

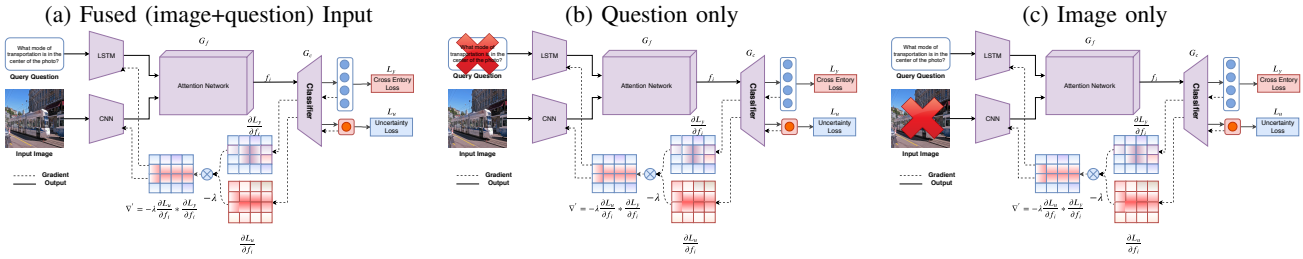

Fig. 6. This figure shows importance of uncertainty on attention map when we consider fused (image and question) as a input , only image as a input and only question as input.

adding a conjugate image pair. It contains 443,757 image-question pairs on the training set, 214,354 pairs on the validation set and 447,793 pairs on the test set, which is more than twice the first version. All the questions and answers pairs are annotated by human annotators. The benchmark results on VQA-v2 dataset is presented in Table-VII.

VQA-HAT [36]: To compare our attention map with human-annotated attention maps, we use VQA-HAT [36] dataset. This dataset is developed for image de-blurring for answering the visual question. It contains 58475 human-annotated attention maps out of 248349 training examples and includes three sets of 1374 human-annotated attention maps out of 121512 validation examples of question image pairs in the validation dataset. This dataset is developed for VQA-v1 only.

VQA-X [57]: We compare our explanation map with human annotated visual explanation maps provided in VQA-X [57] dataset. We calculate rank correlation to compare ground truth explanation map with our explanation, which is present in the main paper in Table-2. VQA-X dataset contains 29459 question answer pair for training set, 1459 pair fro val set and 1968 for test set. Human annotated visual explanation is provided for VQA-v2 only.

B. Ablation Analysis

We provide result of both quantitative and qualitative analysis as in the following sections.

1) Ablation Analysis for Uncertainty Loss: Our proposed GCA model's loss consists of undistorted and distorted loss. The undistorted loss is the Standard Cross-Entropy (SCE) loss. The distorted loss consists of uncertain loss (either aleatoric uncertainty loss (AUL), or predictive uncertainly loss (PUL)),

TABLE III
ABLATION ANALYSIS FOR OPEN-ENDED VQA1.0 ACCURACY ON TEST-DEV

Models	All	Yes/No	Number	Others
Baseline	63.8	82.2	37.3	54.2
VE	64.1	82.3	37.2	54.3
UDL	64.4	82.6	37.2	54.5
AUL	64.7	82.9	37.4	54.6
PUL	64.9	83.0	37.5	54.6
UDL+VE	64.8	82.8	37.4	54.5
AUL+VE	65.0	83.3	37.8	54.7
PUL+ VE	65.3	83.3	37.9	54.9
AUL +UDL	65.6	83.3	37.6	55.0
PUL + UDL	65.9	83.7	37.8	55.2
A-GCA (ours)	66.3	84.2	38.0	55.5
P-GCA (ours)	**66.5**	**84.7**	**38.4**	**55.9**

TABLE IV
RANK CORRELATION FOR EXPLANATION MASK IN VQA-X [57] DATA WITH OUR EXPLANATION MASK USING GRAD-CAM.

Model	RC(↑)	EMD(↓)
Baseline	0.3017	0.3825
Deconv ReLU	0.3198	0.3801
Guided GradCAM	0.3275	0.3781
Aleatoric mask	0.3571	0.3763
Predictive mask	**0.3718**	**0.3714**

Variance Equalizer (VE) loss and Uncertainty Distorted loss (UDL). In the first block of the Table- III, we report the results when these losses are used individually. (Only SCE

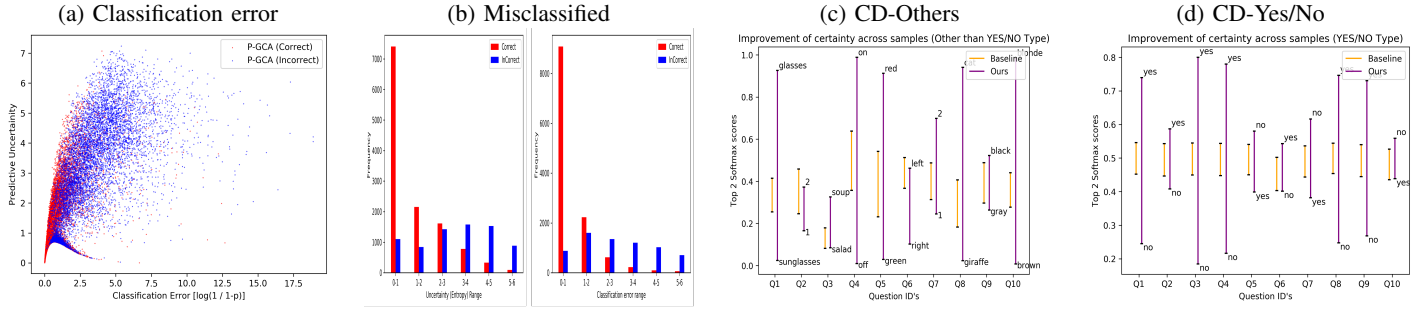

Fig. 7. (a) Uncertainty vs Classification Error plots for our network for 20,000 randomly sampled images. We drew 25 samples of each image using Monte-Carlo sampling from the distribution. (b) Plots showing frequency of samples vs Uncertainty and frequency of samples vs Classification error respectively (c) Distance between the Top 2 Softmax scores for some Questions of type other than *yes/no* (d) Distance between the Top 2 Softmax scores for some Questions of type *yes/no* (Questions corresponding to (c) and (d) could be found in Table- IX and X.)

TABLE V
ABLATION ANALYSIS AND SOTA BETWEEN HAT[36] ATTENTION AND GENERATED ATTENTION MASK

Model	**RC(↑)**	**EMD(↓)**	**CD(↑)**
SAN [36]	0.2432	0.4013	–
CoAtt-W[33]	0.246	–	–
CoAtt-P [33]	0.256	–	–
CoAtt-Q[33]	0.264	–	–
DVQA(K=1)[22]	0.328	–	–
Baseline (MCB)	0.2790	0.3931	–
VE (ours)	0.2832	0.3931	0.1013
UDL (ours)	0.2850	0.3914	0.1229
AUL (ours)	0.2937	0.3867	0.1502
PUL(ours)	0.3012	0.3805	0.1585
PUL + VE (ours)	0.3139	0.3851	0.1631
PUL + UDL(ours)	0.3243	0.3824	0.1630
A-GCA (ours)	0.3311	0.3784	0.1683
P-GCA (ours)	**0.3341**	**0.3721**	**0.1710**
Human [36]	0.623	–	–

TABLE VI
ALEATORIC & EPISTEMIC UNCERTAINTY MEASUREMENT SCORE.

Type of Uncertainty	**Variance**
Aleatoric (with VE)	$3.277 \times 10^{-}3$
Aleatoric (with UDL)	$5.743 \times 10^{-}3$
Aleatoric (with AUL)	$2.561 \times 10^{-}3$
Aleatoric (with PUL)	$1.841 \times 10^{-}3$
Epistemic (50% training)	$4.371 \times 10^{-}4$
Epistemic (75% training)	$3.369 \times 10^{-}4$
Epistemic (100% training)	$1.972 \times 10^{-}4$

loss is there in the Baseline). We use a variant of the MCB [6] model as our baseline method. As seen, PUL, when used individually, outperforms the other 4. This could be attributed to PUL guiding the model to minimize both the data and the model uncertainty. The second block of the Table- III depicts the results when we tried while combining two different individual losses. The model variant, which is guided using the combination of PUL and UDL loss performs best among the five variants. Then finally, after combining (AUL+UDL+VE+SCE), denoting it as A-GCA model and combining (PUL+UDL+VE+SCE), indicating it as P-GCA, we report an improvement of around 2.5% and 2.7% accuracy score respectively.

Further, we plotted Predictive uncertainty (Figure 7(a,b)) of some randomly chosen samples against the Classification error (error=$\log \frac{1}{1-p}$, where p is the probability of misclassification). As seen, when the samples are correct, they are also certain and have less Classification Error (CE). To visualize the direct effect of decreased uncertainty, we plotted (Figure 7(c, d)). It can be seen that how similar classes like (glasses, sunglasses) and (black, gray), etc., thus leading to uncertainty, got separated more in the logit space in the proposed model. Table- IX and - X shows list of the questions and its corresponding id's in present in the Figure 7.

2) Analysis of Epistemic Uncertainty : We measure the uncertainty(entropy) in terms of mean and variance for all the answer prediction in the validation dataset. We also measure uncertainty for an individual question in the dataset. Here, we split our training data into three parts. In the first part, the model is trained with 50% of the training data. In the second part, the model is trained with 75% of training data, and in the third part, the model is trained with the full dataset as shown in the second block of the Table- VI. It is observed that the model uncertainty(epistemic uncertainty) decreases as training data increases.

3) Analysis of Aleatoric Uncertainty : We analyze the contribution of each term in the uncertainty loss to estimate data uncertainty, as shown in the first block of the Table VI. From measurements, it can be seen that comparing aleatoric uncertainty of an image with epistemic uncertainty of another image doesn't make sense because of the significant difference in their values. However, both the uncertainties can be individually compared with the corresponding uncertainties of other images. We capture predictive uncertainty by combining aleatoric uncertainty with the entropy of the prediction (epistemic uncertainty) as mentioned in equation 5 of the main paper. Finally, we provide a variation of aleatoric uncertainty and uncertainty distorted loss over a number of epochs, as shown in Figure 13.

4) Analysis of Attention Maps: We compare attention maps produced by our proposed GCA model, and it's variants with

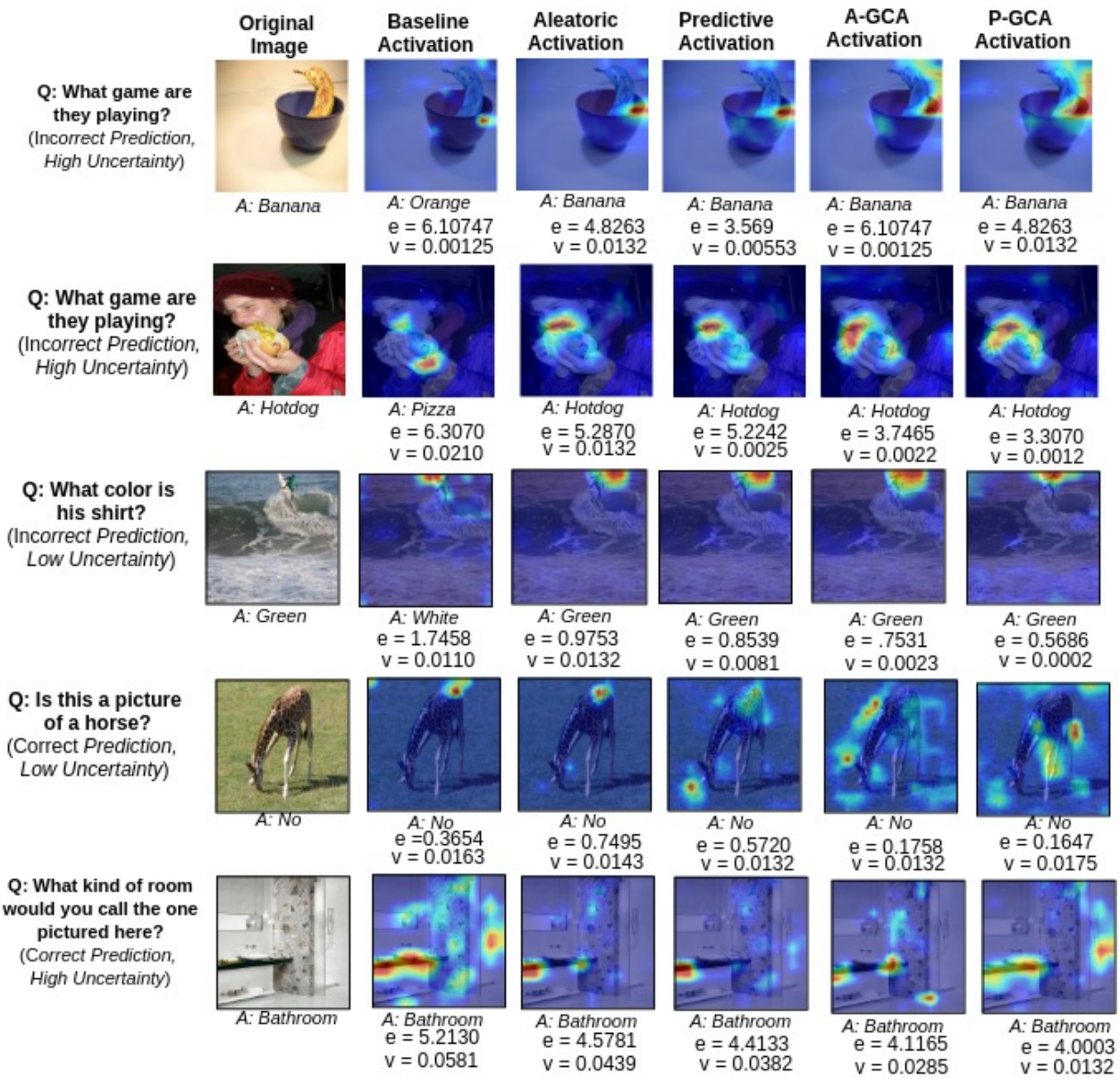

Fig. 8. Examples with different approaches in each column for improving attention using explanation in a self-supervised manner. The first column indicates the given target image and its question and answer. Starting from the second column, it shows the activation map for baseline (MCB) Attention Network, Aleatoric (AUL), Predictive (PUL), A-GCA, P-GCA based approach respectively. The source of original images are refereed from MS-COCO [7] dataset and questions are from VQA [8] dataset.

the base model and reports them in Table-V. Rank correlation and EMD score are calculated for the produced attention map against human-annotated attention (HAT) maps [36]. In the table, as we approach the best-proposed GCA model, Rank correlation (RC) is increasing. EMD is also decreasing (Lower the better) as we move towards GCA. To verify our intuition, that we can learn better attention mask by minimizing the uncertainty present in the attention mask, we start with VE and observe that both rank correlation and answer accuracy increase by 0.42 and 0.3 % from baseline respectively. We also observe that with UDL, AUL, and PUL based loss minimization technique, both RC and EMD improves, as shown in the Table- V. Aleatoric-GCA (A-GCA) improves 5.21% in terms of RC and 2.5% in terms of accuracy. Finally, the proposed Predictive-GCA (P-GCA), which is modeled to consider both data and the model uncertainty improves the RC by 5.51% and accuracy by 2.7% as shown in the Table- V and Table-III. Since HAT maps are only available for VQA-v1 dataset, thus, this ablation analysis has been performed only for VQA-v1. We also providing SOTA results for VQA-v1 and VQA-v2 dataset as shown in Table- VIII and Table- VII respectively. Also, we compare with our gradient certainty explanation with human explanation present in VQA-v2 dataset for the various model as mentioned in Table- IV. This human explanation mask only available for VQA-v2 dataset. We observe that our attention (P-GCA) mask performs better than others as well.

5) Statistical Significance Analysis: We analyze Statistical Significance [58] of our GCA model against the variants

mentioned in section-5.2 of our method as well as other methods for certainty activation map. The Critical Difference (CD) for Nemenyi [59] test depends on given α (confidence level, which is 0.05 in our case) on average ranks and N (number of tested datasets). The low difference in ranks for two models implies that they are significantly less different. Otherwise, they are statistically different. Figure 12 visualizes the post hoc analysis using the CD diagram. It is clear that P-GCA works best and is significantly different from other methods.

6) Analysis on negative results : We have shown this in Figure 9. As is evident, our method tends to fail most often when either the image, question or a combination of both is ambiguous, that is, under a hard setting. In the given examples, we could observe that in first three examples, image information alone does not seem sufficient for reaching to an answer. In the fourth example, the question makes the task difficult, that is, very few data points would have had "hydrant" in the image.

7) When wrong predictions are given by the model, how close/distant are from the true label?: To illustrate this, we show such examples in Figure 10. We could observe that our model tends to struggle with hard examples. But even when our model predicts a wrong answer, the probability distance between the wrong and the correct answer is usually lesser than that in the baseline.

C. Qualitative Result

Visualization of certainty activation maps for image samples as we move from our baseline model to the P-GCA model, as shown in Figure 8. We provide attention map visualization of all models for 5 example images. The first raw, the baseline model misclassifies the answer due to high uncertainty value, that gets resolved by our methods(P-GCA). We can see how attention is improved as we go from our baseline model (MCB) to the proposed Gradient Certainty model (P-GCA). For example, in the first row, MCB is unable to focus on any specific portion of the image, but as we go towards the right, it focuses the cup bottom, (indicated by intense orange color in the map). Same can be seen for other images also. We have visualized Grad-CAM maps to support our hypothesis that Grad-CAM is a very good way for visualizing what the network learns as it can focus on right portions of the image even in the baseline model (MCB), and therefore, can be used as a tutor to improve attention maps. For example, in MCB it tries to focus on the right portions but with the focus to other points as well. However, in our proposed model, visualization improves as the models focuses only on the required portion. We provide attention map visualization for various explanation modules (Guided-Relu, Deconv-Relu) and our (P-GCA) attention map in Figure 11. We also visualise the resulting attention maps of different samples of a particular image-question pair as shown in our project website-1. (Monte Carlo Simulation was done here). (Refer to Equation 4.) From these figures, we observe that the uncertainty is low when certainty map point to the right answer and vice versa.

TABLE VII
SOTA: OPEN-ENDED VQA2.0 ACCURACY ON TEST-DEV

Models	All	Y/N	Num	Oth
SAN-2[21]	56.9	74.1	35.5	44.5
MCB [6]	64.0	78.8	38.3	53.3
Bottom[[60]]	65.3	81.8	44.2	56.0
DVQA[22]	65.9	82.4	43.2	56.8
MLB [35]	66.3	83.6	44.9	56.3
DA-NTN [61]	67.5	84.3	47.1	57.9
Counter[62]	68.0	83.1	51.6	58.9
BAN[24]	**69.5**	85.3	**50.9**	**60.2**
P-GCA + SAN (ours)	59.2	75.7	36.6	46.8
P-GCA + MCB (ours)	65.7	79.6	40.1	54.7
P-GCA + Counter (ours)	69.2	**85.4**	50.1	59.4

TABLE VIII
SOTA: OPEN-ENDED VQA1.0 ACCURACY ON TEST-DEV

Models	All	Y/N	Num	Oth
DPPnet [30]	57.2	80.7	37.2	41.7
SMem[[32]]	58.0	80.9	37.3	43.1
SAN [21]	58.7	79.3	36.6	46.1
DMN[63]	60.3	80.5	36.8	48.3
QRU(2)[38]	60.7	82.3	37.0	47.7
HieCoAtt [33]	61.8	79.7	38.9	51.7
MCB [6]	64.2	82.2	37.7	54.8
MLB [35]	65.0	84.0	37.9	54.7
DVQA[22]	65.4	83.8	38.1	55.2
P-GCA + SAN (ours)	60.4	80.7	36.6	47.9
A-GCA + MCB (ours)	66.3	84.2	38.0	55.5
P-GCA + MCB (ours)	**66.5**	**84.6**	**38.4**	**55.9**

D. Comparison with baseline and state-of-the-art

We obtain the initial comparison with the baselines on the rank correlation on human attention (HAT) dataset [36] that provides human attention while solving for VQA. Between humans, the rank correlation is 62.3%. The comparison of various state-of-the-art methods and baselines are provided in Table V. We use a variant of MCB [6] model as our baseline method. We obtain an improvement of around 5.2% using A-GCA model and 5.51% using P-GCA model in terms of rank correlation with human attention. From this, we justify that our attention map is more similar to human attention map. We also compare with the baselines on the answer accuracy on VQA-v1[8] dataset, as shown in Table- VIII. We obtain an improvement of around 2.7% over the comparable MCB baseline. Our MCB based model A-GCA and P-GCA improves by 0.9% and 1.1% accuracy as compared to state of the art model DVQA [22] on VQA-v1. However, using a saliency-based method [64] that is trained on eye-tracking data to obtain a measure of where people look in a task-independent manner, results in more correlation with human attention (0.49), as noted by [36]. However, this is explicitly trained using human attention and is not task-dependent. In our approach, we aim to obtain a method that can simulate human

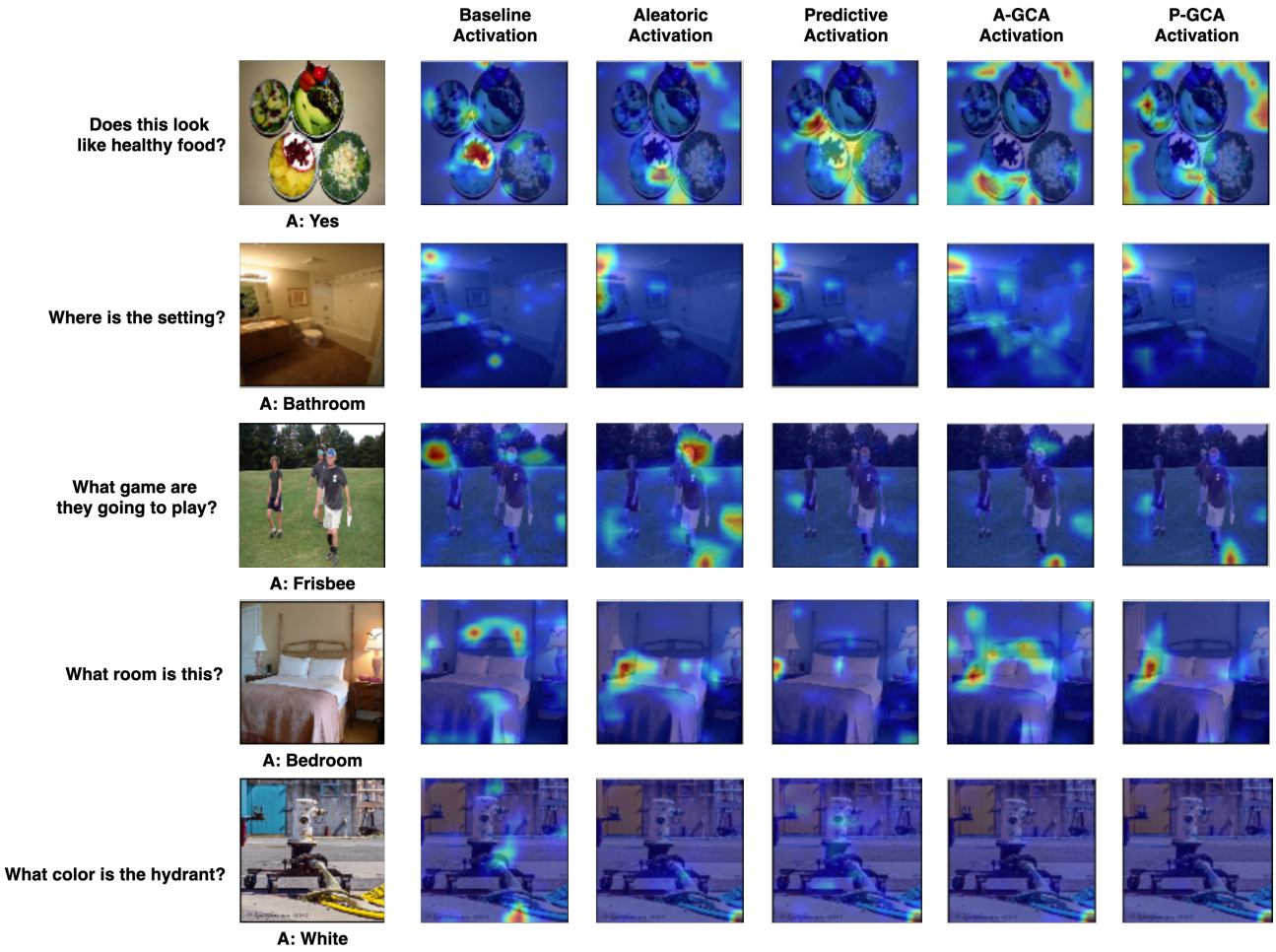

Fig. 9. Some negative results of our model with different approaches in each column. The source of original images are refereed from MS-COCO [7] dataset and questions are from VQA [8] dataset.

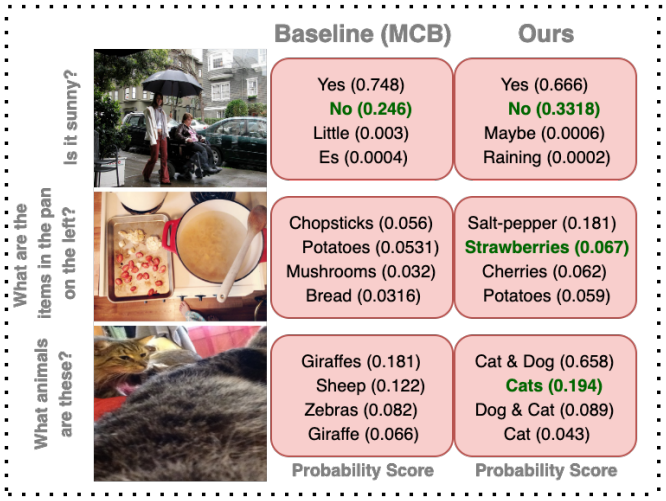

Fig. 10. Top 4 softmax scores for wrongly predicted samples by baseline and our proposed model. (Bold green text highlights the ground truth answer.) The source of original images are refereed from MS-COCO [7] dataset and questions are from VQA [8] dataset.

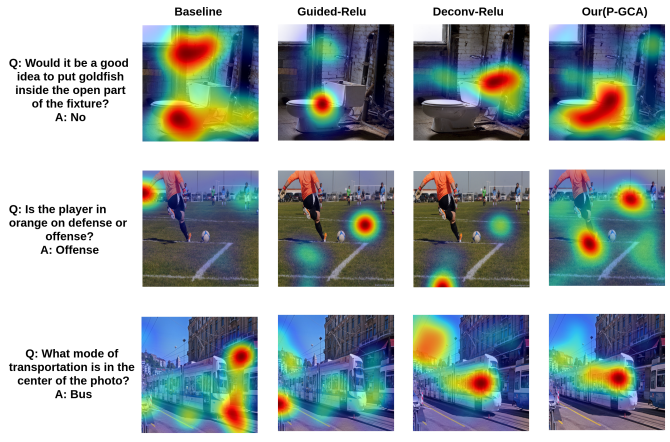

Fig. 11. We compare the attention map of baseline(MCB) Model, Guided-Relu, Deconv-Relu with our (P-GCA) attention map. P-GCA generates better attention maps when compared to other models.

cognitive abilities for solving the tasks. We provide state of the art results for VQA-v2 in Table- VII. This table shows that using GCA method, the VQA result improves. We have provided more results for attention map visualization for both types of uncertainty, training setup, dataset, and evaluation

TABLE IX
REFERENCE FOR THE FIGURE 7(C).

Question	ID
What does the person in this picture have on his face?	1
How many baby elephants are there?	2
What is in the bowl?	3
Is the television on or off?	4
What color is the walk light?	5
Which way is its head turned?	6
How many people are riding on each bike?	7
What animal is in this picture?	8
What color is the road?	9
What color is the boy's hair?	10

TABLE X
REFERENCE FOR THE FIGURE 7(D).

Question	ID
Is this wheat bread?	1
Is the cat looking at the camera?	2
Is this chair broken?	3
Are these animals monitored?	4
Does the cat recognize someone?	5
Is the figurine life size?	6
Is the smaller dog on a leash?	7
Is this in the mountains?	8
Is the woman sitting on the bench?	9
Is the church empty?	10

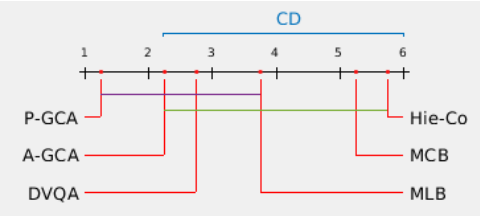

Fig. 12. The mean rank of all the models on the basis of all scores are plotted on the x-axis. CD=3.7702, p=0.003534. Here our P-GCA model and other variants are described in section-5.4. The colored lines between the two models represent that these models are not significantly different from each other.

methods here.[1].

E. Training and Evaluation

1) Model Configuration : We trained the P-GCA model using classification loss and uncertainty loss in an end-to-end manner. We have used ADAM optimizer to update the classification model parameter and configured hyper-parameter values using validation dataset as follows: {learning rate = 0.0001, batch size = 200, beta = 0.95, alpha = 0.99 and epsilon = 1e-8} to train the classification model. We have used SGD optimizer to update the uncertainty model parameter and configured hyper-parameter values using validation dataset as follows: {learning rate = 0.004, batch size = 200, and epsilon = 1e-8} to train the uncertainty model.

2) Evaluation Methods : Our evaluation is based on answer accuracy for VQA dataset and rank correlation for HAT dataset. **Accuracy:** VQA dataset has 3 type of answers: yes/no, number and other. The evaluation is carried out using two test splits, i.e test-dev and test-standard. The question in corresponding test split are of two types: Open-Ended and Multiple-choice. Our model generates a single word answer on the open ended task. For each question there are 10 candidate answer provided with their respective confidence level. This answer can then be evaluated using accuracy metric defined as follows: This answer can be evaluated using accuracy metric provide by Antol *et al.*[8] as follows.

$$Acc = \frac{1}{N}\sum_{i=1}^{N} min\Big(\frac{\sum_{t\in T^i}\mathbb{I}[a_i = t]}{3}, 1\Big) \tag{14}$$

Where a_i the predicted answer and t is the annotated answer in the target answer set T^i of the i^{th} example and $\mathbb{I}[.]$ is the indicator function. The predicted answer ai is correct if at least 3 annotators agree on the predicted answer. If the predicted answer is not correct then the accuracy score depends on the number of annotators that agree on the answer.

Rank Correlation: We used rank correlation technique to evaluate[36] the correlation between human attention map and DAN attention probability. Here we scale down human attention map to 14x14 in order to make same size as DAN attention probability. We then compute rank correlation as mentioned in [22]. Rank correlation technique is used to obtain the degree of association between the data. The value of rank correlation[65] lies between +1 to -1. When R_{Cor} is close to 1, it indicates positive correlation between them, When R_{Cor} is close to -1, it indicates negative correlation between them, and when R_{Cor} is close to 0, it indicates No correlation between them. A higher value of rank correlation is better.

3) Attention Map Visualisation : The size of attention maps is 14×14, and the size of the preprocessed image from COCO-QA is 448×448. To visualize the attention, we need to make attention probability distribution same size as given COCO-QA Image. To do this, we first scale the attention probability distribution to 448×448 using bi-cubic interpolation and then convolve it with a Gaussian filter of size 31×31 with mean 0 and variance 1 to obtain final attention mask. We multiply or mask the obtained attention mask on the original image.

VIII. CONCLUSION

In this paper, we provide a method that uses gradient-based certainty attention regions to obtain improved visual question answering. The proposed method yields improved uncertainty estimates that are correspondingly more certain or

[1]https://delta-lab-iitk.github.io/U-CAM/

(a) Aleatoric Loss (b) Uncertainty Distorted Loss (c) Cross Entropy Loss

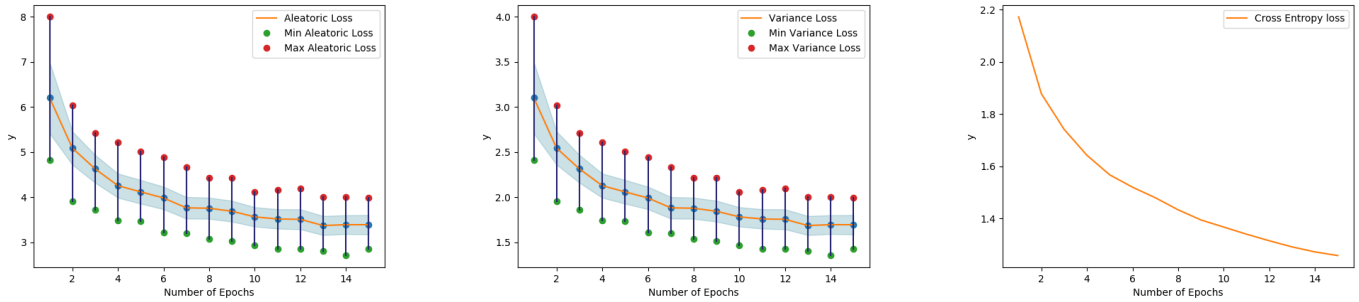

Fig. 13. Variance in Aleatoric Loss, Uncertainty Distorted Loss and Cross Entropy Loss *vs* number of epochs during training. For the first two plots, red and green dots denote the maximum, and the minimum values among all the batches for the corresponding loss, with the blue region showing the spreadness of the loss.

uncertain, show consistent correlation with misclassification and are focused quantitatively on better attention regions as compared to other states of the art methods. The proposed architecture can be easily incorporated in various existing VQA methods as we show by incorporating the method in SAN [21] and MCB [6] models. The proposed technique could be used as a general means for obtaining improved uncertainty and explanation regions for various vision and language tasks, and in future, we aim to evaluate this further for other tasks such as 'Visual Dialog' and image captioning tasks.

REFERENCES

[1] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning (ICML)*, pp. 1050–1059, 2016.

[2] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Advances in neural information processing systems*, pp. 1019–1027, 2016.

[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization.," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *Advances in neural information processing systems*, pp. 5574–5584, 2017.

[5] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," 2018.

[6] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, pp. 740–755, Springer, 2014.

[8] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *International Conference on Computer Vision (ICCV)*, 2015.

[9] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, pp. 818–833, Springer, 2014.

[10] Y. LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*.

[11] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *ICLR (workshop track)*, 2015.

[12] A. Krizhevsky, V. Nair, and G. Hinton, "The cifar-10 dataset," *online: http://www. cs. toronto. edu/kriz/cifar. html*, vol. 55, 2014.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[14] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International Conference on Machine Learning*, pp. 3145–3153, 2017.

[15] P. Kheradpour and M. Kellis, "Systematic discovery and characterization of regulatory motifs in encode tf binding experiments," *Nucleic acids research*, vol. 42, no. 5, pp. 2976–2987, 2014.

[16] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3319–3328, JMLR. org, 2017.

[17] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[18] P. Pasupat and P. Liang, "Compositional semantic parsing on semi-structured tables," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1470–1480, 2015.

[19] M. Everingham and J. Winn, "The pascal visual object classes challenge 2007 (voc2007) development kit,"

[20] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," 2018.

[21] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 21–29, 2016.

[22] B. Patro and V. P. Namboodiri, "Differential attention for visual question answering," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[23] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2017.

[24] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Advances in Neural Information Processing Systems*, pp. 1571–1581, 2018.

[25] B. N. Patro, Anupriy, and V. P. Namboodiri, "Probabilistic framework for solving visual dialog," *Pattern Recognition*, vol. 110, p. 107586, 2021.

[26] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual dialog," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 326–335, 2017.

[27] B. N. Patro, M. Lunayach, S. Patel, and V. P. Namboodiri, "U-cam: Visual explanation using uncertainty based class activation maps," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[28] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[29] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 2953–2961, 2015.

[30] H. Noh, P. Hongsuck Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 30–38, 2016.
[31] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4995–5004, 2016.
[32] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *European Conference on Computer Vision*, pp. 451–466, Springer, 2016.
[33] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances In Neural Information Processing Systems*, pp. 289–297, 2016.
[34] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to compose neural networks for question answering," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2016.
[35] J.-H. Kim, K. W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard Product for Low-rank Bilinear Pooling," in *The 5th International Conference on Learning Representations*, 2017.
[36] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra, "Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.
[37] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4613–4621, 2016.
[38] R. Li and J. Jia, "Visual question answering with question representation update (qru)," in *Advances in Neural Information Processing Systems*, pp. 4655–4663, 2016.
[39] E. H. Shortliffe and B. G. Buchanan, "A model of inexact reasoning in medicine," *Mathematical biosciences*, vol. 23, no. 3-4, pp. 351–379, 1975.
[40] H. C. Lane, M. G. Core, M. van Lent, S. Solomon, and D. Gomboc, "Explainable artificial intelligence for training and tutoring," in *Proceedings of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*, pp. 762–764, 2005.
[41] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros, "What makes paris look like paris?," *ACM Transactions on Graphics*, vol. 31, no. 4, 2012.
[42] T. Berg and P. N. Belhumeur, "How do you tell a blackbird from a crow?," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9–16, 2013.
[43] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.
[44] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013.
[45] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *ICLR (workshop track)*, 2017.
[46] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *International Conference on Machine Learning*, pp. 1613–1622, 2015.
[47] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.
[48] M. Fortunato, C. Blundell, and O. Vinyals, "Bayesian recurrent neural networks," *arXiv preprint arXiv:1704.02798*, 2017.
[49] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
[50] M. Teye, H. Azizpour, and K. Smith, "Bayesian uncertainty estimation for batch normalized deep networks," *arXiv preprint arXiv:1802.06455*, 2018.
[51] L. Smith and Y. Gal, "Understanding measures of uncertainty for adversarial example detection," *arXiv preprint arXiv:1803.08533*, 2018.
[52] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Advances in Neural Information Processing Systems*, pp. 7047–7058, 2018.
[53] V. K. Kurmi, S. Kumar, and V. P. Namboodiri, "Attending to discriminative certainty for domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 491–500, 2019.
[54] T. Alshawi, Z. Long, and G. AlRegib, "Unsupervised uncertainty estimation using spatiotemporal cues in video saliency detection," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2818–2827, 2018.
[55] Y. Gal, *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.
[56] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *stat*, vol. 1050, p. 26, 2017.
[57] D. Huk Park, L. Anne Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8779–8788, 2018.
[58] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.
[59] D. Fišer, T. Erjavec, and N. Ljubešić, "Janes v0. 4: Korpus slovenskih spletnih uporabniških vsebin," *Slovenščina*, vol. 2, no. 4, p. 2, 2016.
[60] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018.
[61] Y. Bai, J. Fu, T. Zhao, and T. Mei, "Deep attention neural tensor network for visual question answering," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 20–35, 2018.
[62] Y. Zhang, J. Hare, and A. Prügel-Bennett, "Learning to count objects in natural images for visual question answering," 2018.
[63] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *Proceedings of International Conference on Machine Learning (ICML)*, 2016.
[64] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Computer Vision, 2009 IEEE 12th international conference on*, pp. 2106–2113, IEEE, 2009.
[65] J. H. McDonald, *Handbook of biological statistics*, vol. 2. 2009.

Badri N. Patro received his PhD degree in Electrical Engineering from Indian Institute of Technology, Kanpur, M.Tech degree in Electrical Engineering from Indian Institute of Technology, Bombay and B.Tech degree in Electronic and telecommunication Engineering from National Institute of Science and Technology, Odisha. He is currently a Postdoctoral researcher Google Research AI, India. Previously he worked as a Lead Engineer at Samsung R&D Institute, Delhi, India, earlier a associate software engineering at Harman International, Pune, India and assistant software engineering at Larsen and Toubro at Mysore. He has served as a reviewer for CVPR, ICCV, ECCV, AAAI, ACL, TIP, PR, BMVC, WACV, and ICVGIP. His research interests are computer vision, natural language processing and machine learning. He is a member of the CVF and ACL.

Mayank Lunayach is currently a Research Fellow at the Wadhwani Institute for Artificial Intelligence in Mumbai, India. He received his B.Tech degree in Electrical Engineering from Indian Institute of Technology, Kanpur. His research interests are computer vision and machine learning. He is a member of the CVF.

Vinay P. Namboodiri received his Mtech and PhD degrees in computer science from Indian Institute of Technology, Bombay. He is currently an associate professor in the Department of Computer Science and Engineering at Indian Institute of Technology, Kanpur, India, and is associated with the DelTA Laboratory for AI and Machine learning Technology, India. He is a Senior Member of IEEE. He was an assistant professor in the Computer Science and Engineering at Indian Institute of Technology, Kanpur. He has served as a reviewer for CVPR, ICCV, AAAI, IJCAI, TIP, BMVC and NAACL. He has served as an area chair for CVPR, ACCV and ICVGIP and served as Senior PC member in AAAI and IJCAI conferences. His research interests are computer vision, natural language processing and machine learning. He is a member of the IEEE and CVF.